Large Language Models for Arabic Automated Essay Scoring

Leena Najjar, Liyakathunisa Syed
Department of Computer Science-College of Computer Science and Engineering, Taibah University, Medina, Saudi Arabia

***Abstract*—Automated Essay Scoring (AES) has become an important research area in educational artificial intelligence due to its potential to support scalable and consistent assessment. The developments within the realm of transformers and large language models (LLMs) have led to great improvements within AESs through their ability to comprehend semantics and context, as well as their knowledge of rubrics. Although these advancements have been realized within the English language, there is still relatively little research surrounding Arabic Automated Essay Scoring (AAES). This survey summarizes some of the latest advancements in AAES and discusses traditional ML models, deep learning, transformers, and LLM-driven evaluation frameworks. In this study, researchers synthesize the relevant literature regarding datasets, prompts, pre-processing methods, performance metrics, and reliability of scores. Typical performance metrics used to analyze the level of agreement between human raters and automated systems are QWK, MAE, and correlation-based metrics. The survey also describes crucial challenges encountered by AAES systems such as insufficient amount of data, inconsistencies, high computation costs, bias, and non-reproducibility. Overall, it can be said that both transformers and LLMs achieve better performance when it comes to capturing context information and providing agreement with human assessment. However, issues with reproducibility and scalability continue to persist. Additionally, the survey presents new areas of research that may be relevant to future AAES studies, such as multilingual evaluation, hybrid grading, explainability, and standardized sources of Arabic essays.**

Keywords*—*Arabic automated essay scoring; large language models; educational AI; prompt engineering; evaluation metrics; natural language processing

I. Introduction

Artificial intelligence is transforming how today's educational systems measure student outcomes, covering aspects ranging from feedback mechanisms to test construction and grading systems [1], [2]. Within this emerging framework, Automated Essay Scoring (AES) is now an important focus of ongoing research, designed to reduce teachers' burdens while increasing consistency, objectivity, and scalability in grading. The earlier Automated Essay Scoring (AES) systems relied on machine learning approaches, which in turn depended on feature engineering. These earlier systems were extremely efficient in the prediction of scores, yet failed to grasp the essence of the essays, the underlying structure of the arguments, and their context [3], [4].

The emergence of deep learning and transformer-based models is an important milestone in the development of automated grading. These models now utilize context and long-range dependencies in essays and other forms of text to better understand language. For instance, Siamese Bi-LSTM models and transformer-based Automated Essay Scoring tools are becoming closer to human judgment than earlier machine-learning-based systems because of their better semantic understanding of the content of the essays [5], [6]. Recently, large language models (LLMs) have moved the AES paradigm towards reasoning-oriented evaluation models that utilize models of natural language interaction, such as rubric-informed prompting, few-shot learning, and instruction tuning to mimic human grading behavior [7], [8].

AAES systems also have their challenges, which are largely due to the complexities of the Arabic language and the lack of sufficiently large datasets for training and evaluating these systems. For instance, although multilingual large language models are highly promising in many evaluation scenarios, their application in Arabic Automated Essay Scoring systems is relatively less explored [8], [9]. In addition, there is a lack of comprehensive systematic studies on the evolution of these systems, dataset characteristics, and evaluation frameworks in the context of Arabic Automated Essay Scoring systems in the existing body of research.

To address this identified information gap, this study aims to synthesize existing research on Arabic automated essay scoring powered by large language models. It discusses the evolution of architectures, the creation of Arabic evaluation datasets, prompt design, the measurement of results, and the ethical issues related to the deployment of AI technology in education. With the synthesis of information from multiple research threads, the study aims to provide a clear reference guide for the development of Arabic automated essay scoring technology that is effective, scalable, and suitable for educational contexts.

The study makes the following contributions to the field of Arabic automated essay scoring powered by large language models: it proposes a comprehensive architectural taxonomy for AAES, presents a discussion on the creation of Arabic evaluation resources, and compares prompting and evaluation techniques for LLM-based essay grading. To the best of our knowledge, this is one of the first systematic surveys on Arabic Automated Essay Scoring (AAES) powered by large language models.

Unlike other surveys on automated essay scoring, this study provides a comprehensive view of Arabic automated essay scoring powered by large language models by discussing together the evolution of architectures, the creation of Arabic evaluation resources, the designing prompts, and the performance evaluation.

The paper is divided into the following sections: Section II discusses the systematic review methodology, Section III synthesizes the findings, Section IV discusses the implications and challenges, and Section V concludes with the future research directions.

II. Methodology

This survey is conducted using a PRISMA-based methodology to maintain transparency, reproducibility, and comparability of the results with other studies on Arabic Automated Essay Scoring (AAES). The relevant papers were retrieved from widely recognized scientific data sources, including IEEE Xplore, ScienceDirect, SpringerLink, MDPI, ACM Digital Library, and arXiv. The scope of the study included papers published between 2018 and 2026, focusing on automated essay scoring, Arabic NLP, large language models, and AI-based educational evaluation.

A four-step paper selection methodology was employed as follows:

- Identification: The relevant papers were identified with the help of the search terms automated essay scoring, Arabic NLP, transformer-based models, and large-language-model-based evaluators.
- Screening: The identified papers were screened by checking the title and abstract to eliminate those irrelevant to the field of education, evaluation, or automated scoring.
- Eligibility: All full papers were checked for eligibility based on set criteria in terms of a distinct methodology being used, technical suitability, and experimental description.
- Inclusion: Papers were considered for final evaluation only if they offered quantitative evaluation metrics like Quadratic Weighted Kappa (QWK), Mean Absolute Error (MAE), correlation coefficients, etc.

Such a method adheres to a well-established and modern blueprint that has been used in recent reviews of AI-powered educational assessment. This methodology strives for a comprehensive and reproducible analysis of how architectures are developed, what kinds of data are used, and how evaluation is conducted in the AAES domain. Before performing the analysis, the review protocol defines clear criteria for categorizing architectures, datasets, and evaluation strategies to ensure consistency across the selected studies.

A. Research Questions

The survey is structured around a set of questions that provide a framework for systematically analyzing how Arabic Automated Essay Scoring (AAES) is developed, what types of data are used, how evaluation is conducted, and the future directions of AAES. These questions guide the selection, comparison, and synthesis of the literature:

RQ1: How do traditional machine learning, deep learning, transformer-based, and large language model approaches compare in terms of performance, reliability, and scoring behavior in Arabic Automated Essay Scoring.

RQ2: What kinds of datasets are being used in Arabic Automated Essay Scoring (AAES) research, and how do linguistic properties, annotation schemes, and domain factors affect model generalization and robustness.

RQ3: What kinds of prompting strategies, scoring systems, and preprocessing methods lead to more consistent, interpretable, and transparent grading in LLM-based evaluation systems.

RQ4: How do evaluation metrics—such as Quadratic Weighted Kappa (QWK), Mean Absolute Error (MAE), and correlation-based metrics—perform in aligning automated grades with human judgments.

RQ5: What kinds of trends in architecture, deployment, and ethics have emerged in LLM-based evaluation systems in Arabic Automated Essay Scoring.

Relevant studies were identified through a systematic search across major academic repositories, including IEEE Xplore, SpringerLink, ScienceDirect, arXiv, as well as peer-reviewed journals on educational artificial intelligence. The search scope was limited to publications between 2018 and 2026, capturing the rapid emergence of transformer architectures and LLM-based grading frameworks in this period.

A comprehensive keyword strategy was employed to ensure wide coverage of Arabic Automated Essay Scoring (AAES) research and related multilingual grading studies. The primary search terms included:

- “Arabic automated essay scoring”
- “large language models grading”
- “LLM evaluation education”
- “transformer essay scoring”
- “rubric-based prompt engineering”
- “multilingual automated grading”
- “QWK automated assessment”

These keywords were applied individually and in combination using Boolean operators to ensure comprehensive retrieval of relevant studies. The search strategy intentionally incorporated both technical, model-oriented terms and education-focused evaluation terminology to capture interdisciplinary contributions spanning natural language processing, educational measurement, and AI-assisted assessment.

A broad inclusion strategy was adopted to encompass experimental studies, methodological frameworks, and conceptual analyses addressing LLM-driven grading pipelines, multilingual evaluation frameworks, and human–AI collaboration in assessment [7], [8]. This approach ensured comprehensive coverage of emerging research trends while remaining aligned with the survey’s research questions and analytical objectives.

An example search string used in IEEE Xplore was:

("Arabic" AND "automated essay scoring") OR ("large language model" AND "educational assessment").

B. Exclusion Criteria (Any of the Following Conditions):

Studies were excluded if they met one or more of the following criteria:

- The study relied on heuristic or rule-based grading approaches without involving machine learning or neural modeling techniques.
- The study lacked experimental validation and/or did not report performance measures.
- The study focused on general natural language processing tasks unrelated to educational assessment and automated grading.

Survey and review articles were included only when they provided essential background information, theoretical underpinnings, and relevant knowledge.

C. Dataset Mapping and Organization

Each included study was carefully reviewed to identify the datasets that were employed in the construction, training, and testing of these models. Special focus was given to Arabic essay datasets such as AR-AES and other emerging Multilingual Automated Essay Scoring benchmarks.

The following aspects were examined to map these datasets:

- Language and linguistic complexity, including the intricacies of the Arabic alphabet and the distinctive morphology of the language.
- Dataset scale and scoring framework, distinguishing between analytic and holistic scoring approaches.
- Domain coverage and educational level, including essays, exams, and other structured writing tasks.
- Annotation quality and inter-rater agreement, measuring the degree of consistency between different raters in evaluating essays.

This dataset-wide mapping enables a structured comparison of architectural performance across different language models, scoring models, and annotation strategies, thus directly addressing Research Question 2. This survey also highlights significant gaps in current Arabic Automated Essay Scoring (AAES) studies, including limited dataset access, domain bias, and scoring rubric inconsistencies, all of which affect model generalization.

D. Architectural Categorization

To address Research Questions 1 and 5 (RQ1 and RQ5), the selected studies were systematically categorized by architectural paradigm. The proposed taxonomy captures the evolutionary trend of Automated Essay Scoring (AES) systems, from traditional feature-engineering models to reasoning-oriented LLM-based evaluators. Four architectural categories have been proposed for the categorization of the studies:

- Traditional Machine Learning AES: Feature-engineering models, including support vector machines, regression models, and traditional large-scale supervised learning models using handcrafted linguistic features and statistical text representations.
- Deep Learning AES: Neural networks combining convolutional and recurrent elements—such as CNN encoders, BiLSTM models, and attention mechanisms—to capture contextual information in essays.
- Transformer-Based Models: Pre-trained transformer models, such as BERT, RoBERTa, and multilingual variants, fine-tuned or task-tuned for essay scoring.
- Large Language Model Evaluators: Prompt-based LLM evaluators using instruction-tuned or generative LLMs for reasoning, interpretation, and evaluation according to educational guidelines.

The proposed architectural categorization of the models provides a unified framework for the analysis of models with different representational power, interpretability, or computational complexity. The evolutionary trend from prediction-based essay scoring toward reasoning-oriented essay evaluation models has been a dominant trend in Arabic AES system research.

Apart from the typical performance-related statistics, this survey is primarily concerned with the more specialized metrics used in Automated Essay Scoring (AES) technology. The aim is to assess the similarity between the results produced by the automated scoring system and those produced by human graders, not merely the proportion of exact matches. The key factors considered for the purpose of this assessment are:

- Quadratic Weighted Kappa (QWK): This compares the results produced by the machine scoring system with the results produced by the human scoring system, taking into consideration the distance or grading gap between the results on the ordinal scale. Because it is more sensitive to the grading gap, this is one of the more popular methods for evaluating the reliability of the scoring system.
- Mean Absolute Error (MAE): This is the average absolute difference between the predicted result and the actual result, providing insight into the accuracy or inaccuracy of the scoring system.
- Correlation Coefficients: Pearson and Spearman correlations are commonly used to evaluate the strength of alignment between automated grading outputs and human judgment patterns, offering complementary insight beyond agreement-based metrics.

Beyond predictive performance, this review also incorporates deployment-oriented evaluation factors, including computational cost, model complexity, prompt stability, and score reproducibility across multiple inference runs. These considerations are particularly relevant for large language model–based grading systems, where stochastic generation processes and high computational requirements may affect

reliability, fairness, and scalability in real-world educational settings.

E. Comparative Analysis

The final step in the review process involved the application of structured synthesis among all the studies selected in order to methodically evaluate architectural performance, dataset impacts, and grading systems. Rather than relying only on the accuracy of the grading systems, the analysis focused on methodological rigor and practical deployment considerations in Arabic Automated Essay Scoring (AAES).

Special emphasis is given to some key areas that affect the reliability of the grading system in using Large Language Model-based grading systems:

- Score stability under stochastic LLM-based inference: It examines how probabilistic generation affects reproducibility across repeated grading runs.
- Effectiveness of rubric-aligned prompting: It assesses how structured evaluation instructions improve agreement with human graders.
- Impact of multi-run scoring and statistical aggregation: It includes averaging and variance-reduction techniques designed to mitigate output variability.
- Trade-offs between interpretability and performance: It is used especially when comparing traditional machine learning models with transformer-based and generative LLM evaluators.

Model capabilities were therefore analyzed not only in terms of predictive performance but also with respect to fairness, transparency, computational feasibility, and deployment practicality in real educational environments. This multidimensional comparison enabled a more nuanced understanding of how architectural design and evaluation methodology jointly influence grading reliability.

Out of 60 studies identified in the initial phase for the candidates, 45 studies proceeded to the title and abstract screening phase after eliminating duplicates and irrelevant literature. Upon checking the full texts, 15 studies were excluded because they lacked quantitative evaluation metrics. Finally, 30 studies were included in the synthesis phase, all of which reported quantitative performance metrics related to automated essay scoring, multilingual AES, or LLM-based educational evaluation. The relatively low number of studies included is due to the fact that Arabic Automatic Essay Scoring is a fairly recent topic of investigation. Most articles excluded from the review failed to provide quantitative evaluation measures and detailed methods.

Our overall framework for synthesizing existing trends and identifying emerging trends in AI-based educational evaluation encompasses literature screening, architectural categorization, dataset analysis, and benchmark evaluation. This framework allows for a transparent and reproducible evaluation of existing and emerging trends in AI-based educational evaluation. Fig. 1 depicts the PRISMA-based flow diagram for identifying existing studies in this survey.

III. Results

In this section, the findings from the literature review are synthesized to provide a cohesive overview of how architectural factors, dataset properties, prompting strategies, and evaluation techniques combine in LLM-based Arabic Automated Essay Scoring (AAES). Rather than simply presenting discrete performance metrics, the findings are organized around the research questions to highlight the development of patterns, the effectiveness of the methods, and the challenges facing reliable grading. The text emphasizes the interplay between architectural, dataset, and prompting factors and evaluation reliability, semantic interpretation, and conformity to human evaluation.

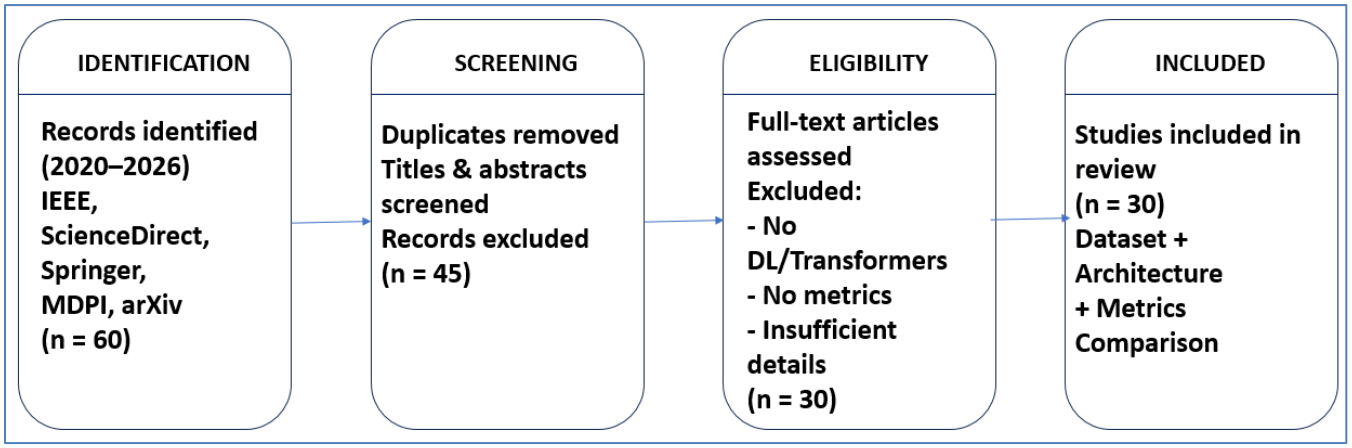

Fig. 1. PRISMA flow diagram of the study selection and analysis process.

A. Performance of Automated Essay Scoring Architectures (RQ1)

Across the surveyed studies, conventional machine learning-based AES systems tend to correlate with human evaluators, albeit at a moderate level, with correlation coefficients ranging generally from 0.55 to 0.70 [3], [4]. These approaches are also popular due to their low computational cost and ease of interpretation, but the difficulty in capturing discourse-level semantic meaning, especially in languages with complex morphology like Arabic, makes it hard to interpret the

significance of linguistic features. Then came the advent of deep learning, especially the BiLSTM and attention-based models, which improve contextual interpretation through learning hierarchical representations. These models tend to achieve improved Quadratic Weighted Kappa (QWK) scores compared to traditional machine learning approaches, reflecting better sensitivity to coherence, argument structures, and lexical variation within essays [5], [6]. Thus, the introduction of deep learning methods into the AES task marked a significant improvement over conventional machine learning methods, which relied on the engineering of linguistic features.

The introduction of Transformer-based models also took the AES task to the next level, especially with the introduction of bidirectional contextual embeddings, which capture long-range dependencies within essays. Indeed, several studies report higher QWK performance compared to earlier approaches, highlighting the effectiveness of self-attention mechanisms in capturing semantic relationships within essays [6], [9].

Recently, the advent of large language model-driven grading pipelines points to the emergence of reasoning-oriented evaluators rather than purely scoring-oriented ones. For instance, prompt-based frameworks facilitate the interpretation of grading rubrics, the generation of structured evaluations, and the provision of feedback according to the objectives.

The literature indicates that it is possible to train evaluators, such as large language models, to be as effective as human evaluators [7], [10]. However, a challenge remains: inference variability, where different evaluations can produce different results [10]. Statistical methods are needed to ensure reproducibility. Table I provides a summary of the differences, trends, and limitations of the AES architectures, including the shift from feature-oriented to reasoning-oriented evaluators, such as the LLM.

B. Dataset Characteristics and Their Influence on Model Behavior (RQ2)

Dataset characteristics play a significant role in the performance and generalization of Arabic Automated Essay Scoring (AAES) systems. In the reviewed works, factors such as dataset size, linguistic diversity, annotation practices, and scoring mechanisms are identified as the primary dataset characteristics that affect model performance and generalization.

The AR-AES dataset [9] is part of the rare set of structured datasets designed specifically for Arabic essay scoring, which supports both analytic and holistic essay scoring mechanisms. Its relatively controlled linguistic environment is advantageous for training and experimentation. This limited diversity of subjects and writing styles may affect the generalization of the AAES system, especially when used outside the original context or with different learners. On the other hand, multilingual corpora like ASAP-AES [11] are commonly employed for Automated Essay Scoring since they are large in size and include a scoring framework. However, because these corpora focus mainly on English essays, they are not directly applicable to Arabic AES due to the unique linguistic and rhetorical features. New datasets, such as ZaQQ [12], show potential for addressing shortcomings of existing datasets by combining human and AI annotation. Although these datasets aim to improve evaluation by including diverse linguistic perspectives, they are still in early development. As a result, many LLM-based grading systems are sensitive to data composition and may fail outside their original training domain on which they were initially trained.

One of the commonalities in the reviewed works is the importance of dataset diversity and annotation standards in LLM-based grading systems. The studies have shown how these systems perform poorly when faced with different writing styles, educational backgrounds, and task formats, especially when the dataset used to train the LLM is narrow in domain. Thus, increasing the size and standardization of Arabic essay datasets is essential to improve evaluation.

Table II provides a comparative analysis of the major datasets used in the research on Arabic AES, including their linguistic coverage, applications, and limitations.

C. Impact of Prompting Strategies and Scoring Frameworks (RQ3)

Prompt engineering has emerged as a key factor influencing the reliability, transparency, and reproducibility of LLM-based Automated Essay Scoring systems. Across the reviewed studies, rubric-aligned prompting consistently improved grading reliability by guiding models to evaluate essays using explicit pedagogical criteria rather than relying solely on implicit semantic similarity [7]. Structured prompts that explicitly reference scoring rubrics were shown to enhance alignment with human evaluators and reduce ambiguity in generated assessments.

Recent research also highlights the effectiveness of bilingual prompting strategies, in which Arabic evaluation instructions are complemented by English rubric descriptors. This approach leverages the strong multilingual capabilities of contemporary LLMs, improving interpretability and supporting more stable alignment with human grading standards, especially in multilingual educational contexts [8].

A notable methodological advancement is the adoption of multi-run scoring frameworks, which address the stochastic nature of generative model inference [10]. By producing multiple grading outputs and aggregating them using statistical techniques—such as averaging, variance reduction, or consensus scoring—researchers mitigate output instability and improve agreement metrics, including Quadratic Weighted Kappa (QWK) and Mean Absolute Error (MAE). These findings suggest that grading reliability is influenced not only by model architecture but also by the design of the prompting and aggregation pipeline.

Additional prompting strategies reported in the literature include:

- Few-shot prompting, where curated examples illustrate grading expectations and provide contextual guidance.
- Structured reasoning prompts, encouraging step-by-step evaluation aligned with educational rubrics and promoting explainability.

- Error-detection prompts, designed to identify inconsistencies or logical conflicts in generated scoring outputs.

Collectively, these techniques demonstrate that prompt design functions as a methodological layer comparable in importance to model architecture, directly affecting grading stability, interpretability, and fairness. Explainability is particularly important in educational assessment because grading decisions may directly influence student progression, feedback quality, and instructor trust in automated evaluation systems. Table III summarizes the primary prompting and preprocessing strategies employed in LLM-based AES systems and highlights their reported contributions to scoring reliability.

D. Evaluation Metrics and Alignment with Human Graders (RQ4)

Evaluation in Arabic Automated Essay Scoring (AAES) goes beyond conventional accuracy-based metrics to capture the nuanced relationship between automated predictions and human judgment. Across the reviewed studies, Quadratic Weighted Kappa (QWK) remains the most widely adopted evaluation measure, as it accounts for ordinal score differences and quantifies agreement between automated grading outputs and human annotations [6]. Complementary metrics, such as Mean Absolute Error (MAE), provide an interpretable measure of scoring deviation, while correlation coefficients—including Pearson and Spearman—assess linear alignment between model predictions and human grading.

LLM-based evaluators demonstrate promising performance across these evaluation criteria, particularly when integrated with rubric-driven prompting strategies and multi-run aggregation frameworks. These methodological enhancements improve agreement scores and reduce prediction variance. However, the literature emphasizes that strong numerical performance does not necessarily guarantee fairness, interpretability, or pedagogical validity. Several studies report cases in which models achieve high correlation or QWK values while producing inconsistent reasoning processes or exhibiting bias toward specific linguistic styles, essay lengths, or structural patterns [7], [8].

As a result, recent research increasingly advocates for complementary evaluation paradigms that incorporate qualitative analysis of model-generated feedback, human-centered evaluation protocols, and transparency-oriented assessment frameworks. These approaches aim to bridge the gap between quantitative agreement metrics and broader educational objectives of automated assessment systems.

A comparative overview of computational and evaluation trade-offs across AES architectures is presented in Table IV, highlighting how differences in model complexity, prompting strategies, and inference mechanisms influence both grading performance and practical deployment considerations.

In order to provide a quantitative analysis of architectural performance trends in the reviewed AES systems, Table V presents an overview of the evaluation measures, performance metrics, and trade-offs encountered by each of the leading architectures in AES.

According to Table V, AES based on transformers and large language models demonstrates superior agreement with human graders compared to machine-learning-based AES systems. Nevertheless, they are characterized by higher computational complexity and instability of grading.

E. Emerging Trends and Opportunities in Arabic AES (RQ5)

The literature review revealed several emerging trends that are changing how Arabic automated essay scoring with AI technology is carried out. Overall, there is a shift from traditional prediction-based grading systems to approaches that focus more on reasoning and human-centered evaluation.

The first trend is the shift towards reasoning-based grading systems that use large language models. Compared to traditional feature-based or semantic grading systems, large language models better understand context and provide more accurate evaluations for essays in different languages [7], [8].

The second trend is the focus on human-AI collaborative grading systems. Compared to AI-only grading systems, recent approaches focus on human-AI collaborative systems where large language models assist in grading essays and provide explanations for their decisions [10].

The third trend is the focus on explainability, fairness, and responsible use in Arabic automated essay grading systems. Compared to AI-only grading systems for Arabic essays, recent approaches emphasize prompt transparency and bias detection in large language models.

The fourth trend is the focus on developing multilingual foundation models for Arabic automated essay grading systems. Compared to LLM-only grading systems for Arabic essays, recent approaches aim to develop more robust systems capable of adapting to different educational contexts and environments [9].

The assembled evidence suggests that there is a paradigm shift in the field of automated essay scoring, from machine learning pipelines based on feature engineering to reasoning-oriented evaluation systems facilitated by the use of transformer models and large language models. This shift is part of the larger paradigm transition in the field of educational AI, from predictive scoring models to more semantically grounded models of evaluation that facilitate reasoning and the production of structured feedback based on rubrics. While the transformer and LLM models have shown clear improvements in contextual understanding and human-aligned evaluation, they have also led to the identification of methodological and deployment issues related to reproducibility in the presence of stochasticity, computational demands, and the identification of issues related to the ethics of deployment as core concerns in the deployment of Arabic Automated Essay Scoring systems.

F. Balancing Linguistic Understanding and Grading Reliability

A notable feature of the studies examined here is the intrinsic trade-off between the semantic reasoning capacity and the scoring stability of the automated essay scoring systems. Conventional machine learning techniques, such as support vector machines and regression-based grading models, offer good interpretability and efficiency; however, the feature

engineering limitations of such systems hinder the accurate modeling of discourse semantics and contextual relationships, especially for morphologically rich languages such as Arabic [3], [13]. Hence, while such systems frequently demonstrate scoring stability, the outputs may not generalize well for assessing higher-order writing competencies such as argumentation quality, coherence, and semantics.

The adoption of deep learning and transformer-based models has greatly improved the contextual modeling capacity of the essay scoring systems. Pre-trained models like AraBERT and multilingual transformers improve the model's capacity to grasp the meaning of words, which improves the alignment of the automated scoring system with the scoring of human graders [6], [9]. However, it is essential to note that the models rely on the quality of the datasets. This is because the datasets for the Arabic education domain are limited. This implies that the model's performance in an experimental setting does not necessarily reflect the real-world situation.

Large language models (LLMs) offer a further advancement for automated essay scoring by supporting reasoning-based evaluation rather than purely predictive scoring. When guided by well-designed prompts, LLMs can interpret scoring criteria, generate structured explanations, and provide pedagogically meaningful feedback [7], [10]. Nevertheless, because these models operate probabilistically, repeated evaluations may produce variable results, raising concerns about consistency and reproducibility in grading.

TABLE I. COMPARISON OF AUTOMATED ESSAY SCORING ARCHITECTURES

Ref	Model category	Example architectures	Performance trend	Primary application	Limitations
[2], [14]	Traditional machine learning	SVM, regression-based AES	Moderate agreement with human graders	Educational analytics	Limited contextual understanding
[3], [4]	Deep learning models	BiLSTM, CNN–LSTM	Improved contextual evaluation	Structured essay grading	Requires large labeled datasets
[6], [9]	Transformer-based models	BERT, AraBERT, multilingual transformers	Strong semantic representation	Multilingual and Arabic AES	Computationally expensive
[15]	Hybrid architectures	Multi-model fusion, ensemble learning	Stable grading performance	Cross-dataset AES evaluation	Complex training pipelines
[16], [17]	Large language models	GPT-based and instruction-tuned LLMs	High reasoning and alignment potential	Rubric-guided scoring	Output variability

TABLE II. DATASETS USED IN ARABIC AND CROSS-LINGUAL AES RESEARCH

Ref	Dataset	Language	Task type	Typical usage	Identified issues
[12]	ZaQQ dataset	Arabic	Human–AI collaborative AES	LLM evaluation research	Limited scale
[13], [18]	ASAP-AES benchmark	English	Holistic essay scoring	Model benchmarking	Not Arabic-native
[19]	Arabic essay corpus	Arabic	Grammar-based scoring	Educational assessment	Annotation inconsistency
[9]	Arabic grammatical AES	Arabic	Linguistic evaluation	Transformer-based grading	Domain-specific bias
[20]	Multilingual AES datasets	Mixed languages	Cross-lingual grading	LLM evaluation studies	Domain shift

TABLE III. PROMPTING AND PREPROCESSING STRATEGIES

Ref	Technique	Reported improvement	Affected models	Observations
[7], [21]	Rubric-guided prompting	Improved human alignment	GPT-style graders	Enhances grading consistency
[10]	Few-shot prompting	Reduced scoring variance	Instruction-tuned LLMs	Provides contextual guidance
[22], [23]	Multi-run evaluation	Stabilized grading outputs	LLM evaluators	Mitigates stochastic behavior
[9]	Arabic text normalization	Better semantic accuracy	Transformer-based AES	Handles morphology and dialects
[24], [25]	Structured reasoning prompts	Improved transparency	Generative AI graders	Supports ethical AI evaluation

TABLE IV. COMPUTATIONAL TRADE-OFFS ACROSS AES ARCHITECTURES

Ref	Model	Approximate parameters	Computational cost	Evaluation metrics	Deployment suitability
[3], [13]	SVM-based AES	Low	Low CPU usage	MAE, correlation	High for schools
[6]	Transformer AES (BERT)	~110M	Moderate GPU usage	QWK, F1-score	Medium
[9]	Arabic transformer AES	~120M	Moderate	QWK, accuracy	Medium
[7], [26]	GPT-style LLM grading	Billions	Very high cloud cost	Human alignment metrics	Low for edge deployment
[20], [27]	Distilled LLM AES	Reduced parameters	Medium	QWK, MAE	High future potential

TABLE V. QUANTITATIVE PERFORMANCE TRENDS ACROSS AES ARCHITECTURES

Ref	Architecture	Typical Metrics	Performance Trend	Main Limitation
[3], [4]	Traditional ML	MAE, Correlation	Moderate agreement	Weak semantic understanding
[5], [6]	Deep Learning	QWK, MAE	Improved contextual scoring	Requires large datasets
[6], [9]	Transformer-based	QWK, Accuracy	Strong semantic alignment	High computation
[7], [10]	LLM-based AES	QWK, Human alignment	Strong reasoning capability	Output variability

G. Dataset Limitations and Generalization Challenges

One of the difficulties associated with AAES studies involves the scarcity of Arabic essay corpora. Although corpus-based models of English essays, such as the ASAP-AES dataset, help compare the performance of various grading algorithms in AAES, the nature of the texts in those corpora differs greatly from Arabic writing. AAES, too, faces many NLP challenges associated with the Arabic language, such as morphological richness, spelling/orthography variation, the presence or absence of diacritics, and regional dialectal variations. This presents a problem for transformer-based and LLM-based AAES, which are typically built using only MSA-based and multilingual corpora. Moreover, the current Arabic educational corpora used in AAES research are not sufficiently large or diverse [6], [9].

As the literature surveyed in the previous sections shows, the way a system behaves is heavily influenced by the evaluation resource it was exposed to. Inconsistencies in the way the resources are labeled may cause difficulties in reproducing results from one paper to another. Domain shifts, such as the one between essay prompts and actual student responses to these prompts, also cause difficulties in generalizing the results. In these cases, the performance tends to degrade as the system is exposed to new contexts.

Nevertheless, the recent advances in large language models are alleviating some of the difficulties researchers face in the field. Large language models generalize well to other semantic spaces and enhance the performance of the system on cross-lingual and cross-genre tasks [8]. However, the lack of good resources for the Arabic essay domain still poses a significant barrier to the evaluation process.

H. Influence of Prompting Strategies and Preprocessing

Prompt design is currently at the center of the effectiveness and reliability of automated essay scoring models that use large language models. Current research shows that prompt-based methods that align with a rubric and use thoughtfully designed few-shot examples allow models to better align their scores with human raters since the model is designed to evaluate the student response according to a specific framework. By integrating the scoring rubric into the prompt, these prompt-based methods force the model to evaluate the student response more thoughtfully rather than simply relying on language generation. When combined with other prompt-based methods that incorporate reasoning-based prompts, the process is more transparent since the model will outline the steps it takes to evaluate the student response. This is in line with new AI ethics that focus on accountability and transparency [25].

However, it is also worth noting that preprocessing methods that take into account the nuances of the Arabic language also greatly improve the accuracy of the semantic understanding and grading. By taking into account issues such as orthographic normalization and morphology-based tokenization, the model is better able to filter out irrelevant information that arises from variations in spelling and morphology. This is particularly important for models that use a transformer-based architecture and a large language model since token-level information is particularly important for the final grading. However, prompt-based methods are still a source of variability for LLM-based AES models. The fact that small changes to the prompt and scoring rubric can result in a change to the final score also brings into question the reproducibility and fairness of the grading process. Recent studies also highlight risks related to hallucinated feedback, prompt sensitivity, and bias amplification, where slight prompt variations or linguistic patterns may influence grading consistency and fairness across student populations.

I. Computational Cost and Deployment Feasibility

This is particularly the case despite the marked improvements in terms of performance that have been reported in the literature on the use of LLMs in AES systems. For example, the use of external inference services is not transparent in terms of the exact ways in which the models function, and it also limits the extent of control that can be exercised by the educational institutions in question, which is particularly problematic in terms of governance and other issues of relevance.

From the point of view of computational complexity, classical machine learning models in the AES system remain attractive, despite their limitations in terms of their ability to reason and their low computational requirements [3]. On the other hand, the Transformer model offers contextual evaluation efficiently, without the computational costs that have been reported in the application of LLMs in the AES system.

Despite the clear improvements in the application of LLMs in the AES system, the literature indicates that there are high computational requirements, even when cloud computing and/or hardware accelerators are used [7], [26]. Moreover, the challenges associated with the use of external APIs have led to the development of techniques for efficient fine-tuning, reducing the computational requirements without compromising the performance in terms of the ability to reason.

The literature indicates the need to balance the expressiveness of the model and its practicality, especially in the application of LLMs in the AES system, in terms of the ability to reason and the computational requirements.

J. Emerging Trends in AI-Based Educational Assessment

The results of this survey indicate that there are several trends that are currently shaping the way in which educational assessments with AI will be carried out in the future. One of the trends that has been identified in this regard is that of hybrid assessments that use a combination of large language model-based reasoning with other scoring mechanisms. Instead of purely relying on the output of the language model, hybrid assessments use rule-based metrics, statistical features, and rubric-based prompts to make grading easier to understand. By using this approach, hybrid assessments have been shown to eliminate randomness in grading without losing the flexibility of language model-based grading.

Another trend that has been identified in this regard is that of multilingual and multi-cultural assessments. As educational assessments with AI move beyond purely English-based datasets, there is an emerging need to create grading models that take into account linguistic diversity, dialectical diversity, and educational diversity. Multilingual assessments focus on adapting grading criteria to specific educational contexts within countries while maintaining reliability across languages [11]. This trend is in keeping with the need to create globally accessible AI-based assessments.

Ethics-based governance is also an emerging trend in this regard. The Artificial Intelligence Assessment Scale has been developed to create guidelines that encourage explainability, bias detection, and responsible AI-based education [24], [25].

Recent discussions also emphasize fairness across Arabic dialects, mitigation of cultural bias in grading, and protection of student educational data privacy during large-scale AI-assisted assessment. This is particularly important in grading contexts since grading decisions often determine student progression. As such, new research in this field has shown that there is an emerging trend of human-centered assessments that use explainability, auditability, and pedagogical oversight in language model-based grading pipelines.

K. Implications for Future Arabic AES Research

The survey indicates several areas that should be pursued to further advance Arabic Automated Essay Scoring, so that it can eventually work well in classrooms. First off, there is a need to develop larger and more diverse Arabic essay datasets. Current datasets are small, not diverse in topics, and the annotation quality is inconsistent. Future datasets must include standardized scoring rubrics, multi-rater annotation, and cover diverse settings to allow reproducible benchmarking.

Another important research direction involves robust prompting methods and evaluation benchmarks to stabilize Arabic AES with large language models. Prompts are known to significantly affect generative model outputs, so standardized rubric-aligned prompting frameworks should be developed, methods to aggregate results over several runs, and transparent reporting. Developing shared evaluation methods will allow us to better compare architectures, which will help unify the field.

The third area that should be pursued is multimodal inputs in Arabic AES. Multimodal inputs, which include other signals beyond text, such as handwriting, student interaction, or curriculum information, could enable Arabic AES to assess more than text quality to better understand other factors. This is in line with other areas of educational AI, which are moving toward more holistic evaluation methods, not just text-based evaluation.

The last area that should be pursued is addressing practical challenges, beyond peak performance metrics. While most studies focus on higher QWK or correlation scores, few studies investigate other important factors, such as computational cost, data privacy, scalability of infrastructure, and usability of Arabic AES. These factors are important to transition Arabic AES from experimental settings to practical and deployable educational technology.

For future research on Arabic AES, retrieval-augmented grading framework, specialized Arabic LLMs in addition to parameter-efficient fine-tuning methods tailored to educational purposes can be adopted. Moreover, collaborative evaluation models that involve humans in the loop may offer better clarity in grading while still enabling human control over the process.

L. Limitations of the Survey

This review aggregates the current progress in Arabic Automated Essay Scoring (AAES) systems, but there are several caveats that should be noted. The process of selecting the relevant literature heavily relied on indexed digital libraries, which may introduce publication bias or omit newer regional literature. The findings in the literature have been based on reported outcomes and not a unified benchmark, which makes it difficult to compare the literature due to differences in the datasets used, the rubrics for scoring, and the methods of evaluation. The field of large language models is also evolving at a rapid rate, which further limits the temporal relevance of the literature, with newer models and methods being developed frequently. The lack of large standardized datasets of essays in Arabic is also a problem that limits the generalization of the literature.

IV. Conclusion

This review distills the latest advances in Arabic Automated Essay Scoring to demonstrate how new developments in transformer-based models and large language models are revolutionizing the field of automated education assessment. The results of the research demonstrate that recent advances in natural language processing have revolutionized the field of automated grading to an unprecedented level of fidelity in determining semantic coherence, language quality, and grading rubric alignment. While the older approach to automated grading relied on the use of various features to arrive at an assessment of the student response to an assignment, the new approach of using transformer-based models and large language models has clearly ushered in a new era of grading that is more context-oriented and reasoning-based.

However, the results of the research also demonstrate that while there has been an unprecedented level of progress in the field of AAES, there remains much to be addressed in terms of the complexity of the task. One of the problems is that there is currently a bottleneck in terms of the amount of data available in the Arabic language to support AAES. While domain variety in available datasets is limited to support AAES, there is also an inconsistency in the annotations of the available data. While

multilingual pretraining can be used to overcome some of these problems to an extent, language diversity remains a critical factor in determining the robustness of grading.

The survey also shines a light on how prompting strategies and preprocessing pipelines are at the core of how effective an evaluation carried out by an LLM-based evaluator turns out to be. While strategies such as rubric-aware prompting, structured reasoning prompts, and Arabic normalization have proven to be effective in improving grading clarity and consistency, however, these generative models remain highly sensitive to prompting strategies. This is an area that calls for standardized prompting strategies to be developed. However, there is no doubt that ethical considerations have now become an important part of how AI-based evaluators contribute to the educational field. This is because of the broader developments in the direction of developing trustworthy AI in education.

From an implementational perspective, a balance must be maintained between evaluator effectiveness and deployment practicality of the model in an educational setting. While traditional ML approaches and lightweight transformer-based models can be attractive for large-scale educational settings because of their practicality, ease of interpretability, and predictability, there is also the possibility that an LLM-based evaluator can be more effective in its reasoning and feedback mechanisms. However, there is also the flip side of how computationally intensive they can be to deploy, with potential variability in scores due to model stochasticity. Recent developments in parameter-efficient fine-tuning and distillation-based approaches suggest promising avenues to achieve high performance while maintaining deployability.

Looking forward into the future, it is possible to highlight a few research areas that are considered crucial for advancing Arabic automated essay scoring. For example, creating large standardized datasets for essays written in Arabic with standardized annotation schemes will be instrumental in advancing robustness and enabling comparative evaluations across different models. Another direction that may minimize random variability in evaluation outcomes while maintaining interpretability is the use of hybrid grading models that combine symbolic evaluation with generative reasoning. Finally, integrating multimodal educational signals, including handwriting recognition, student metadata, and broader learning analytics, may be very beneficial in advancing evaluation accuracy and personalization.

To conclude, the shift toward LLM-based evaluation for evaluation in place of more traditional automated evaluation systems represents a significant leap forward in AI-based evaluation in educational settings. The current approaches already point toward significant scalability and support for teachers in evaluation processes. However, future research should emphasize robustness, fairness, and interpretability in advancing Arabic automated essay scoring in order to ensure that it moves beyond experimental systems toward more robust systems that align with pedagogical needs and support modernization in evaluation processes.

REFERENCES

[1] H. Niemi, R. D. Pea, and Y. Lu, "Introduction to AI in Learning: Designing the Future," in AI in Learning: Designing the Future, Springer, 2022, pp. 1–15. doi: 10.1007/978-3-031-09687-7_1.

[2] D. S. Yadav, "Navigating the Landscape of AI Integration in Education: Opportunities, Challenges, and Ethical Considerations for Harnessing the Potential of Artificial Intelligence for Teaching and Learning," BSSS Journal of Computer, vol. 15, no. 1, pp. 38–48, 2024, doi: 10.51767/jc1503.

[3] D. S. V. Madala, A. Gangal, S. Krishna, A. Goyal, and A. Sureka, "An Empirical Analysis of Machine Learning Models for Automated Essay Grading." 2018. doi: 10.7287/peerj.preprints.3518v1.

[4] V. V. Ramalingam, A. Pandian, P. Chetry, and H. Nigam, "Automated Essay Grading Using Machine Learning Algorithm," Journal of Physics: Conference Series, 2018, doi: 10.1088/1742-6596/1000/1/012030.

[5] G. Liang, B. W. On, D. Jeong, H. C. Kim, and G. S. Choi, "Automated Essay Scoring: A Siamese Bidirectional LSTM Neural Network Architecture," Symmetry, vol. 10, no. 12, 2018, doi: 10.3390/sym10120682.

[6] S. Ludwig, C. Mayer, C. Hansen, K. Eilers, and S. Brandt, "Automated Essay Scoring Using Transformer Models," Psych, vol. 3, no. 4, pp. 897–915, 2021, doi: 10.3390/psych3040056.

[7] A. Gandolfi, "GPT-4 in Education: Evaluating Aptness, Reliability, and Loss of Coherence in Solving Calculus Problems and Grading Submissions," International Journal of Artificial Intelligence in Education, vol. 35, no. 1, pp. 367–397, 2025, doi: 10.1007/s40593-024-00403-3.

[8] I. D. Mienye, N. Jere, G. Obaido, O. O. Ogunruku, E. Esenogho, and C. Modisane, "Large Language Models: An Overview of Foundational Architectures, Recent Trends, and a New Taxonomy," Springer Nature, 2025, doi: 10.1007/s42452-025-07668-w.

[9] S. Mahmoud, E. Nabil, and M. Torki, "Automatic Scoring of Arabic Essays: A Parameter-Efficient Approach for Grammatical Assessment," IEEE Access, vol. 12, pp. 142555–142568, 2024, doi: 10.1109/ACCESS.2024.3470728.

[10] A. Kundu and D. Barbosa, "Are Large Language Models Good Essay Graders?" Sep. 2024. [Online]. Available: http://arxiv.org/abs/2409.13120

[11] Z. Ke and V. Ng, "Automated Essay Scoring: A Survey of the State of the Art," in Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19), International Joint Conferences on Artificial Intelligence Organization, 2019, pp. 6300–6308. doi: 10.24963/ijcai.2019/879.

[12] Y. Elsayed, E. Nabil, M. Torki, S. Faizullah, and A. Khalafallah, "ZaQQ: A New Arabic Dataset for Automatic Essay Scoring via a Novel Human–AI Collaborative Framework," Data, vol. 10, no. 9, 2025, doi: 10.3390/data10090148.

[13] A. Singh, "Automated Essay Scoring Using Machine Learning," International Journal of Advance Research, Ideas and Innovations in Technology (IJARIIT), 2019, [Online]. Available: https://www.ijariit.com

[14] T. Zhao, "AI in Educational Technology." 2023. doi: 10.20944/preprints202311.0106.v1.

[15] M. Uto, I. Aomi, E. Tsutsumi, and M. Ueno, "Integration of Prediction Scores from Various Automated Essay Scoring Models Using Item Response Theory," IEEE Transactions on Learning Technologies, vol. 16, no. 6, pp. 983–1000, 2023, doi: 10.1109/TLT.2023.3253215.

[16] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing," ACM Computing Surveys, vol. 55, no. 9, pp. 1–35, 2023, doi: 10.1145/3560815.

[17] OpenAI, "GPT-4 Technical Report," arXiv preprint arXiv:2303.08774, 2023, [Online]. Available: https://arxiv.org/abs/2303.08774

[18] H. Li, C. H. Chen, K. Fan, C. Young-Johnson, S. Lim, and Y. Feng, "Agreement Between Large Language Models and Human Raters in Essay Scoring: A Research Synthesis," arXiv preprint arXiv:2512.14561, 2025, [Online]. Available: https://arxiv.org/abs/2512.14561

[19] S. Esmaile, O. Al-Awaida, and B. Alshargabi, "Automated Arabic Essay Grading System based on Support Vector Machine and Text Similarity Algorithm," International Journal of Advanced Computer Science and Applications, 2019, [Online]. Available: https://meu.edu.jo/libraryTheses/5d3c0808d27b7_1.pdf

[20] R. F. Mello et al., "Automatic Short Answer Grading in the LLM Era: Does GPT-4 with Prompt Engineering beat Traditional Models?," in Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK '25), ACM, 2025, pp. 93–103. doi: 10.1145/3706468.3706481.

[21] C. Grévisse, "LLM-based automatic short answer grading in undergraduate medical education," BMC Medical Education, vol. 24, no. 1, p. 1060, 2024, doi: 10.1186/s12909-024-06026-5.

[22] R. Ghazawi and E. Simpson, "Automated Essay Scoring in Arabic: A Dataset and Analysis of a BERT-based System," arXiv preprint arXiv:2407.11212, 2024, [Online]. Available: https://arxiv.org/abs/2407.11212

[23] W. Mansour, S. Albatarni, S. Eltanbouly, and T. Elsayed, "Can Large Language Models Automatically Score Proficiency of Written Essays?" Apr. 2024. [Online]. Available: http://arxiv.org/abs/2403.06149

[24] E. D. Lindsay, M. Zhang, A. Johri, and J. Bjerva, "The Responsible Development of Automated Student Feedback with Generative AI," in EDUCON, 2025. doi: 10.1109/EDUCON62633.2025.11016572.

[25] M. Perkins, L. Furze, J. Roe, and J. MacVaugh, "The Artificial Intelligence Assessment Scale (AIAS): A Framework for Ethical Integration of Generative AI in Educational Assessment," Journal of University Teaching and Learning Practice, vol. 21, no. 6, 2024, doi: 10.53761/q3azde36.

[26] Z. I. Karsa and B. Goldschmidt, "Automatic Evaluation of Programming Tasks Supported by Language Models," IEEE Access, vol. 13, pp. 147741–147756, 2025, doi: 10.1109/ACCESS.2025.3601448.

[27] K. Ono and A. Morita, "Evaluating Large Language Models: ChatGPT-4, Mistral 8x7B, and Google Gemini Benchmarked Against MMLU," Mar. 2024, doi: 10.36227/techrxiv.170956672.21573677/v1.